Collaborative Transformers With Multi-Level Forensic Attention for Image Manipulation Localization

Jiwei Zhang[1,4], **Wenbo Feng**[1], **Siwei Wang**[2*], **Feifei Kou**[1*], **Haoyang Yu**[3], **Shaozhang Niu**[1]

[1]School of Computer Science (National Pilot School of Software Engineering), BUPT, Beijing, China
[2]The Intelligent Game and Decision Lab, Academy of Military Sciences, Beijing, China
[3]China Mobile Internet Co., Ltd, GuangZhou, China
[4]Key Laboratory of Interactive Technology and Experience System, Ministry of Culture and Tourism(BUPT), Beijing, China

Abstract

The proliferation of the tampered images on social media can pose serious societal risks, influencing public opinion and causing panic. Image Manipulation Localization technique has advanced to address this, but some methods focus on microscopic traces, overlooking macroscopic semantics that deceive viewers. To address this problem, we propose a novel Image Manipulation Localization framework called Collaborative Transformers (Co-Transformers), designed to fully explore and utilize the collaborative information between macroscopic semantics and microscopic traces. This framework is based on two Vision Transformer variants. The first variant captures the semantic logic of the image. The second variant delves into microscopic tampering traces. By dynamically fusing these two complementary features, the framework enables interaction between macroscopic semantic inconsistencies and microscopic abnormal traces, effectively coordinating their relationship in the latent space. Furthermore, we introduce a new Multi-Level Forensic Attention (MLF-Attention) mechanism to enhance the model's ability to extract various tampered traces, this mechanism can be integrated into our framework. Compared with existing methods, our proposed framework achieves state-of-the-art results in localization accuracy and shows good robustness against various attacks.

Introduction

In recent years, the rapid development of image processing technology and widespread use of editing software have significantly lowered the barrier to image tampering, leading to the increasingly widespread dissemination of tampered images on social media platforms. As shown in Figure 1, the primary image tampering techniques include splicing, copy-move, and removal. However, the improper use of these techniques, particularly in creating fake news and spreading rumors, can seriously disrupt social order and attract broad public attention. Thus, it is essential to effectively detect and accurately locate these tampered images.

Unlike traditional semantic segmentation, image manipulation localization typically focuses on inconsistencies in non-semantic information, often manifested as discrepancies at tampered object boundaries. Existing methods can be classified into single localization structures and hybrid localization structures. Single localization structures (Wang et al. 2022b; Triaridis and Mezaris 2024) struggle to simultaneously model macroscopic semantic inconsistencies and microscopic tampering traces. Hybrid architectures (Li et al. 2024; Zhang et al. 2023; Zhu et al. 2025) usually combine CNNs, Transformers, or other structures with an additional connector module to bridge features between components. This connector design implicitly reflects a fundamental latent space misalignment between CNN and Transformer representations, hindering effective feature fusion. In our approach, we avoid this issue by employing two specialized Transformers, each selectively extracting distinct forgery-related cues. Instead of relying on complex fusion modules, we adopt a simple strategy: concatenating the outputs of the two Transformers along the channel dimension and feeding the combined features directly into the prediction module.

Furthermore, certain methods using conventional attention mechanisms for Image Manipulation Localization (IML) are constrained by a singular receptive field. For pixel-level anomalies, an overly large receptive field may cause anomalous features to be overlooked amid adjacent normal regions, reducing detection sensitivity. For regional-level distribution discrepancies, a small receptive field limits the model's ability to capture broader semantic con-

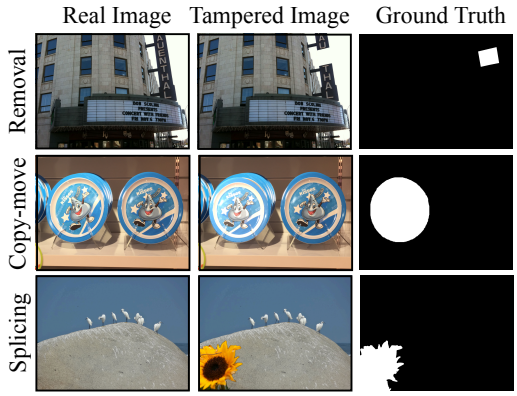

Figure 1: Here are three image tampering techniques: Splicing, Copy-move and Removal.

*Corresponding authors.

text, impairing accuracy in identifying such inconsistencies. This makes it difficult for models to simultaneously leverage both pixel-level anomalies and regional-level distribution discrepancies effectively.

To address the above challenges, we propose a novel framework named Collaborative Transformers. This framework analyzes the features of tampered images using a dual-path collaborative strategy that integrates multi-scale forensic evidence. Then, macroscopic semantic inconsistencies and microscopic forensic anomalies are mutually reinforced via a dynamic fusion module. Furthermore, we introduce Multi-Level Forensic Attention (MLF-Attention) to overcome the limitations of classical attention mechanisms (Vaswani et al. 2017) in IML tasks. By decoupling the feature map into multiple levels and constructing attention mechanisms with varying receptive fields, the model can effectively capture multi-source tampering traces, ranging from pixel-level anomalies to regional-level distribution inconsistencies. These design substantially enhances the accuracy of IML tasks.

Our main contributions are:

1. We propose a novel Collaborative Transformers (Co-Transformers) framework. This framework employs a collaborative mechanism to achieve complementary enhancement of multi-scale tampering cues. The dual-path processing part effectively identifies both microscopic tampering traces, such as edge artifacts and noise distribution anomalies, and macroscopic semantic inconsistencies. Tampering trace across diverse frequency bands is adaptively integrated by a multi-scale feature fusion module, thereby enabling the model to achieve precise localization.

2. We introduce a Multi-Level Forensic Attention (MLF-Attention) mechanism. This mechanism employs a feature separation operation to generate multiple, complementary feature streams. Each stream focuses on representing microscopic tampering traces within distinct receptive field scales, thereby achieving comprehensive utilization of multi-level tampered cues.

3. Evaluations on publicly available benchmark datasets demonstrate that our method achieves state-of-the-art performance, with significant improvements in accuracy, robustness, and generalization.

Related Work

Single localization structures

With the rapid development of deep learning technology, Image Manipulation Localization (IML) methods based on CNNs have made significant progress. PSCC-Net (Liu et al. 2022) features a dual-path mechanism. Its bottom-up path employs progressive, multi-scale mask refinement, aided by dense cross-scale connections, to achieve accurate coarse-to-fine localization of tampered regions. CAAA(Qu et al. 2024) addresses the data scarcity issue by proposing a paradigm to automatically construct a large-scale dataset from web images, and based on it, develops the APSCNet model for accurate manipulation localization.

As the Vision Transformer (ViT)(Dosovitskiy et al. 2021) architecture has shown significant advantages in computer vision tasks such as semantic segmentation, gradually becoming a mainstream method, researchers have begun to explore its application potential in IML tasks. IML-ViT (Ma et al. 2024a) is the first work to introduce a pre-trained ViT model to the IML task. Fine-tuning the model using only the CASIAv2 dataset (Dong, Wang, and Tan 2013), not only solves the problem of data scarcity in IML tasks but also significantly reduces model training costs.

Hybrid localization structures

Hybrid methods combining CNNs and Transformers have also attracted widespread attention. ObjectFormer (Wang et al. 2022a), an architecture that incorporates EfficientNet (Tan and Le 2019) as its initial encoder stage, downsampling the input data into specific feature blocks before passing them to the ViT. Mesorch (Zhu et al. 2025) is a mesoscopic representation-based image manipulation localization method that fuses macroscopic semantics and microscopic trace features through a parallel Transformer-CNN architecture, achieving collaborative analysis of cross-scale features. CNN-T GAN (Zhang et al. 2023) employs a generator to produce a mask resembling the ground truth, and a discriminator to differentiate between real and generated image pairs. UnionFormer (Li et al. 2024) introduces a unified learning framework that integrates tampering clues from three complementary views: RGB features, noise artifacts, and object-level inconsistency. This method proposes a Boundary Sensitive Feature Interaction Network, where a Feature Coupling Unit combines CNN and Transformer outputs.

Method

As shown in Figure 2, our proposed framework consists of three modules: a Noise Fusion Module, Collaborative Transformers, and a Prediction Module for decoding the image. The Noise Fusion Module is responsible for extracting noise features from the image. Following the processing of the noise features and the image by the Collaborative Transformers,, the mask is generated by the Prediction Module. In addition, we further design a Multi-Level Forensic Attention to model the hidden information between noise patches after different noise fusions. All of the details are described below.

Collaborative Transformers

Define the input image as $X \in R^{H\times W\times 3}$. The image X is simultaneously processed by two ViT variants: the Hierarchical Edge-Supervised Transformer (HES-TRansformer) and the Cross-trace Extraction Transformer (CE-Transformer) . For the latter, the images are first pre-processed by the Noise Fusion Module. These modules are utilized to extract edge features and high-frequency noise features of the tampered region, respectively.

Hierarchical Edge-Supervised Transformer adopts the hierarchical encoder architecture of SegFormer (Xie et al. 2021) as its encoder backbone, which employs a hierarchical structure to extract features from images, outputting a series of results at different scales:

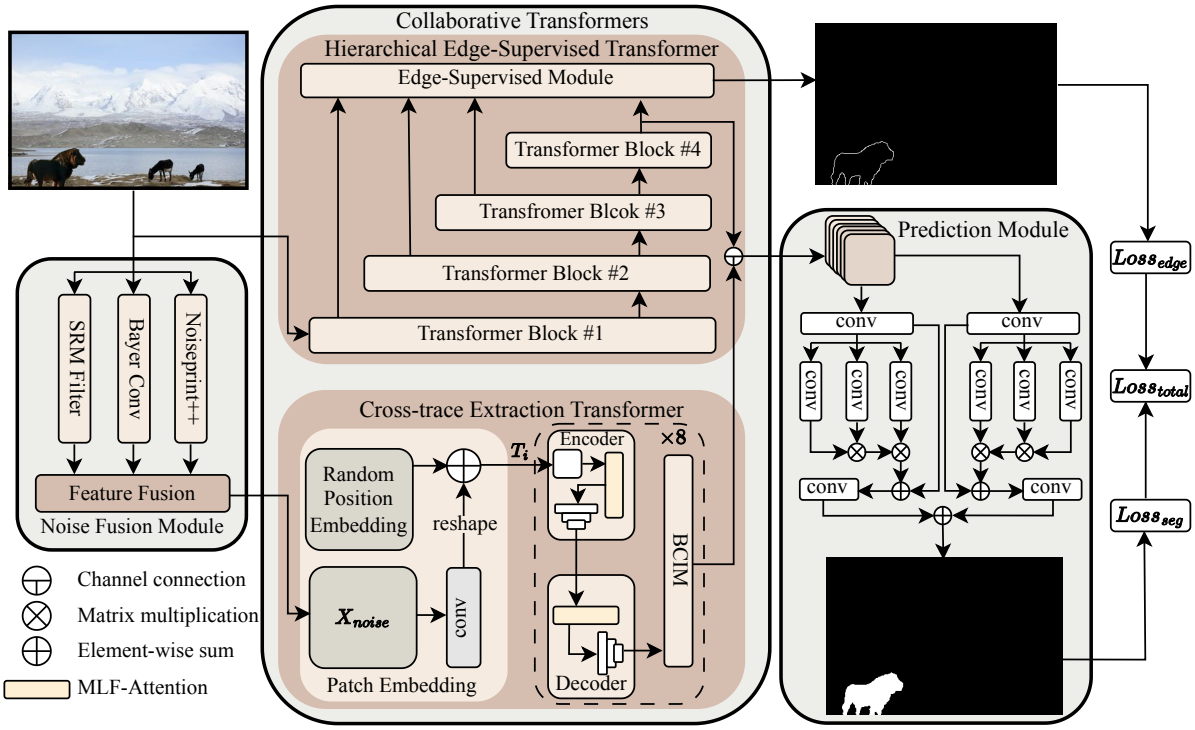

Figure 2: **The framework of the proposed Collaborative Transformers.** High-frequency noise features are first extracted and fused. Then, an Hierarchical Edge-supervised Transformer extracts multi-scale edge features, while a Cross-trace Transformer decouples noise features to capture tampering clues. Finally, all features are dynamically aggregated to produce a pixel-level localization result.

$$\{F_1, F_2, F_3, F_{edge}\} = \text{SegFormer}(X)$$
$$F_i \in R^{\frac{H}{2^{(i+1)}} \times \frac{W}{2^{(i+1)}} \times C_{i_{edge}}} \quad (1)$$

$C_{i_{edge}}$ represents the total number of channels output by the SegFormer at each layer. A staged feature extraction strategy is employed by this encoder. This strategy allows for the rich color information from the RGB streams to be retained and multi-scale features to be extracted through progressive downsampling. Feature maps encompassing varying resolutions are generated by each layer, thereby enabling the capture of both microscopic details and macroscopic semantics.

The decoder part adopts the Edge-supervised Module (Dong et al. 2022), used to locate the boundary regions of the tampered image and calculate the edge loss.

$$X_{edge} = \text{ESM}(F_1, F_2, F_3, F_{edge}) \quad (2)$$

Rich contextual information is provided to the Edge-supervised Module by this multi-scale feature, thereby overcoming the problem of limited local receptive fields caused by the fixed convolution kernels of traditional CNNs.

Cross-trace Extraction Transformer is used to extract the fused noise information $X_{noise} \in R^{H \times W \times 3}$. As shown in Figure 2, the noise information X_{noise} is mapped to the latent space through a convolution operation to generate the corresponding feature representation. Subsequently, the embeddings are obtained from the convolution operation. They are then element-wise added to randomly generated positional encodings to incorporate spatial position information. Then, the embeddings, represented by T_i in the figure, are input into multiple transformer blocks.

The implementation details of the transformer blocks are shown in Figure 3. Each block consists of an encoder and a decoder. T_i first undergoes feature interaction through layer normalization and MLF-Attention. Then, feature mapping is performed through a series of linear layers and normalization layers, using GeLU (Hendrycks and Gimpel 2016) as the activation function. Finally, the features are passed to the next layer through residual connections and linear projection. It is expressed by the following formulas:

$$A_{oe} = \text{MLF-Attention}(q_i, k_i, v_i)$$
$$m_i = (q_i \oplus A_{oe}) \oplus \mathscr{G}(q_i \oplus A_{oe}) \quad (3)$$

where q_i represents the layer-normalized output of Gaussian noise, and k_i and v_i represent the layer-normalized outputs of T_i, and $\mathscr{G}$ represents the linear layers and normalization layers used for feature mapping. Using Gaussian noise as the query provides the data with a diverse initial distribution, helping the model explore the latent representations of different objects in the early stages of training, and uniformly covering the feature space, allowing it to more comprehensively establish associations with different parts of the

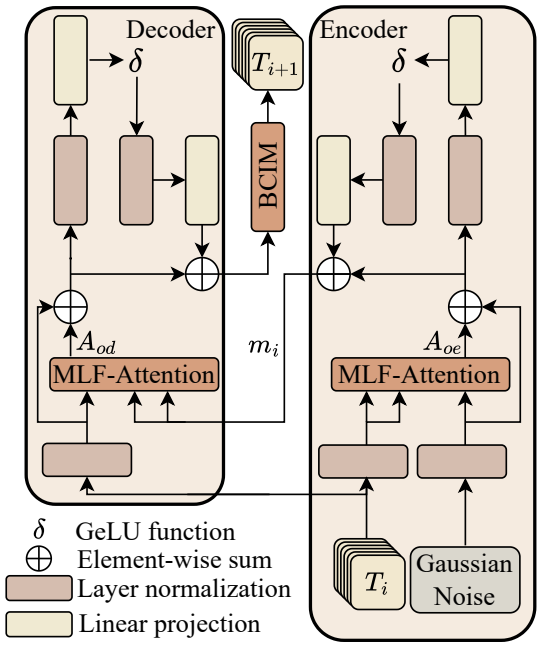

Figure 3: Block of the Cross-trace Extraction Transformer. From T_i to T_{i+1}, the input is the output of the previous block, and the output is the input of the next block.

fused noise.

The decoder part is similar to the encoder part, only the input of the MLF-Attention is different. Then, the Boundary sensitive Contextual Incoherence Modeling (BCIM) (Wang et al. 2022a) is used to finely detect the abnormal area. The formulas are as follows:

$$\begin{aligned} A_{od} &= \text{MLF-Attention}(q_i, m_i, m_i) \\ T_{i+1} &= \text{BCIM}((q_i \oplus A_{od}) \oplus \mathscr{G}(q_i \oplus A_{od})) \end{aligned} \tag{4}$$

q_i represents the layer-normalized output of T_i, and m_i is the mid-level feature.

By stacking 8 transformer blocks as described above, the feature representation can be gradually and deeply extracted and optimized, ultimately obtaining the fused noise feature.

To better integrate these two feature streams, we adopt a prediction module with a dynamic fusion strategy (Fu et al. 2019), enabling the model to adaptively integrate different tampering features and achieve accurate localization of the tampered region. Specifically, the Hierarchical Edge-Supervised Transformer effectively captures the boundary features, while the Cross-trace Extraction Transformer extracts the high-frequency noise features, both contributing to the final localization.

$$X_{loc} = \text{PredictionModule}(F_{edge}, F_{noise}) \tag{5}$$

Multi-Level Forensic Attention

To fully utilize the information after the fusion of multiple noise features, we propose MLF-Attention to capture new traces generated by the fusion between noises. Unlike the classic attention mechanism (Vaswani et al. 2017), we split

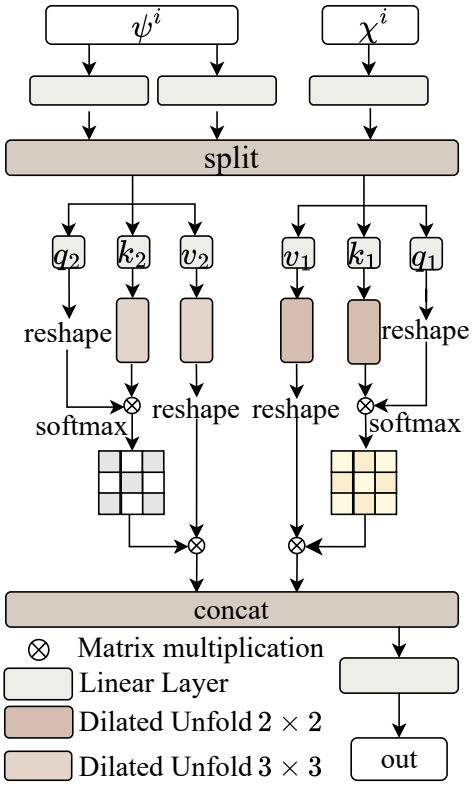

Figure 4: The detailed calculation process of Multi-Level Forensic Attention.

the input features, which can dynamically adjust the receptive field range of features at different levels, greatly promoting collaborative learning of related regions in the same image, and suppressing the influence of irrelevant regions during the feature learning stage.

As shown in Figure 4, for the input features χ and ψ, we first perform feature transformation through linear projection and hierarchical processing. Subsequently, we apply an unfold operation on the features extracted from ψ, flattening them into multiple vectors containing local features, and construct Q_i, K_i, V_i required for the attention mechanism through a concatenation operation. This process can be expressed as the following formulas:

$$\begin{aligned} Q_i &= \text{Reshape}((\text{Split}(W_Q \chi))) \\ K_i &= \text{Unfold}((\text{Split}(W_K \psi))) \\ V_i &= \text{Unfold}((\text{Split}(W_V \psi))) \end{aligned} \tag{6}$$

where, $i \in \{1,2\}$, W_Q, W_K and W_V represent the weight matrices of Q_i,K_i and V_i respectively.

We construct a multi-level analysis mechanism by setting the dilation rates to 2 and 3 for the local unfold operation. The receptive field with a smaller dilation rate can effectively capture microscopic noise anomaly patterns, thereby achieving precise localization of subtle forgery traces. The receptive field with a larger dilation rate covers macroscale noise correlation features comprehensively by modeling long-range dependencies. This multi-scale design main-

tains the fineness of local features while also effectively integrating contextual information of multiple noises.

Besides, introducing local unfold operations with different dilation rates helps enhance the model's robustness when facing different attack. When an image is attacked by Gaussian noise, the continuity of local textures and edges is often disrupted. Traditional attention mechanisms, due to their fixed local receptive fields, are prone to misjudgments of tampered regions. In contrast, through the layered unfold operation, noise features of different scales can be captured more effectively, thereby significantly improving the model's robustness to noise.

Then, attention matrix calculation is performed on two multi-head self-attentions, respectively. By sparsely selecting keys and values centered on the query, the locality and sparsity characteristics of the traditional ViT are preserved. In the tampered boundary area, different attention heads can focus on feature dimensions such as hidden information of the shooting camera, texture continuity, and color consistency, effectively amplifying the subtle differences between the tampered areas and real areas in the feature space. It effectively aggregates multi-scale semantic information within the concerned receptive field and efficiently reduces the self-attention mechanism with computational costs.

After the multi-head self-attention calculation, we obtain two attention matrices of different scales, namely F_1 and F_2 . We concatenate these two attention matrices along the channel dimension and perform feature fusion through a fully connected layer to obtain the final attention output F_{out}.

$$F_{out} = \text{Linear}(\text{Concat}(F_1, F_2)) \tag{7}$$

Loss Function

We construct our loss function on two scales. At the pixel scale, this improves the model's sensitivity to pixel-level tampering localization, and at the edge scale, it learns features unrelated to semantics. Considering that tampered regions generally occupy a small proportion of the original image, we choose to use Dice loss (Milletari, Navab, and Ahmadi 2016) as both the pixel-level loss function and the edge-level loss function. Its definition can be given as follows:

$$loss_{dice}(x) = 1 - \frac{2\sum_{i,j} S(x_{i,j}) \cdot y_{i,i}}{\sum_{i,j} S^2(x_{i,j}) + \sum_{i,j} y_{i,j}^2}, \tag{8}$$

where $y_{i,j} \in \{0,1\}$ is a binary label indicating whether pixel (i, j) is tampered with, and $p_{i,j}$ represents the probability.

Since contextual information and edge tampering detection can enhance each other, we fully exploit their potential complementary relationship. We define the loss function for the optimization process as follow:

$$L_{total} = \alpha L_{edge} + \beta L_{pixel} \tag{9}$$

where L_{edge} and L_{pixel} represent the edge loss and pixel loss, respectively, α and β re the weight coefficients of the two losses. We set α is 0.8 and β is 0.16, thus forcing the model to pay more attention to edge information.

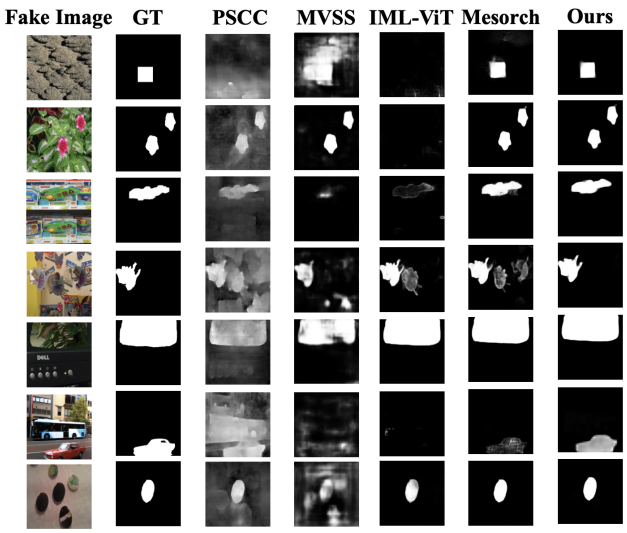

Figure 5: Comprehensive performance evaluation for image manipulation detection and localization methods.

Experiment

Training Data and Implementation Details

Our model is trained using the standard Protocol-CAT(Ma et al. 2024b). This protocol applies various typical data augmentation techniques and resizes all images to a resolution of 512 $\times$ 512 pixels. The training process is conducted on a single NVIDIA 4090 GPU for a total of 200 epochs, with a batch size of 4. The learning rate uses a cosine scheduling algorithm (Loshchilov and Hutter 2017), gradually decreasing from an initial value of 1e-4 to 5e-7, and gradually warming up the learning rate in the first 2 epochs. We use the AdamW optimizer and set the weight decay to 0.05 to mitigate the problem of model overfitting. In addition, by setting the accumulated gradient to 4, we further adjust the batch size, thereby enhancing the model's localization ability on different data.

Test Dataset and Evaluation Metric

We conduct a comprehensive evaluation of the model using widely recognized benchmarks. Specifically, we test the model on the following five datasets: CASIAv1, Coverage, NIST16, Columbia, and AutoSplice. These datasets are invaluable for assessing the generalization capabilities of image manipulation localization methods as they encompass a variety of tampering types and are widely acknowledged by the research community.

To ensure a fair comparison, our method has been evaluated on publicly recognized benchmarks in line with common practice. We report both the standard pixel-level F1-score and the permute F1-score, calculated at a threshold of 0.5, to provide a comprehensive assessment of localization accuracy. Furthermore, we have ensured that the other selected open-source methods utilize the same training procedures as ours to guarantee a fair comparison.

Model	F1-score / Permute F1-score					
	Coverage	Columbia	NIST16	CASIAv1	AutoSplice	Avg.
MVSS-Net(Dong et al. 2022)	0.5188 / 0.5503	0.7322 / 0.7772	0.2876 / 0.3377	0.5486 / 0.5707	0.3779 / 0.5570	0.4894 / 0.5586
PSCC-Net(Liu et al. 2022)	0.3791 / 0.4415	0.8640 / 0.8924	0.3681 / 0.4143	0.5472 / 0.5583	**0.5531** / 0.6831	0.5423 / 0.5987
IML-ViT(Ma et al. 2024a)	0.5364 / 0.5852	0.9334 / 0.9721	0.1100 / 0.1503	0.7392 / 0.7512	0.2687 / 0.5676	0.5184 / 0.6053
CAT-Net(Kwon et al. 2022)	0.4272 / 0.5165	0.9151 / 0.9547	0.3787 / 0.3316	0.8140 / 0.8154	0.3870 / 0.6168	0.5844 / 0.6470
TruFor(Guillaro et al. 2023)	0.4573 / 0.5369	0.8845 / 0.9547	0.3480 / 0.4046	0.8176 / 0.8340	0.3830 / **0.6910**	0.5781 / 0.6843
Mesorch(Zhu et al. 2025)	0.5862 / 0.6342	0.8903 / 0.9708	0.3921 / 0.4514	**0.8398 / 0.8472**	0.4004 / 0.6709	0.6138 / 0.7149
Co-Transformers (ours)	**0.6344 / 0.6770**	**0.9412 / 0.9863**	**0.4256 / 0.4698**	0.8074 / 0.8198	0.4276 / 0.6863	**0.6473 / 0.7278**

Table 1: Model performance comparison using standard F1 and permute-F1 metrics(Kwon et al. 2022). Best scores are in bold and the suboptimal scores are underlined.

Model	F1-score / Permute F1-score					
	Coverage	Columbia	NIST16	CASIAv1	AutoSplice	Avg.
w/o noiseprint++	0.6341 / 0.6774	0.9287 / 0.9834	0.4401 / 0.4826	0.8096 / 0.8244	0.4231 / 0.6952	0.6471 / 0.7326
w/o bayer	0.5940 / 0.6470	0.9374 / 0.9838	0.4137 / 0.4594	0.8079 / 0.8202	0.4805 / 0.6986	0.6437 / 0.7218
w/o srm	0.6653 / 0.7056	0.9341 / 0.9757	0.3915 / 0.4404	0.8027 / 0.8202	0.4254 / 0.6818	0.6438 / 0.7173
with noiseprint++	0.6145 / 0.6621	0.9251 / 0.9899	0.4457 / 0.4891	0.7980 / 0.8139	0.4147 / 0.6912	0.6369 / 0.7292
with bayar	0.6406 / 0.6678	0.9418 / 0.9650	0.4102 / 0.4672	0.7877 / 0.8029	0.4582 / 0.6976	0.6457 / 0.7161
with srm	0.6097 / 0.6482	0.9393 / 0.9717	0.4041 / 0.4557	0.8206 / 0.8323	0.4254 / 0.6891	0.6398 / 0.7194
full feature extractors	0.6344 / 0.6770	0.9412 / 0.9863	0.4256 / 0.4698	0.8074 / 0.8198	0.4276 / 0.6863	0.6473 / 0.7278

Table 2: An ablation analysis of Co-Transformer performance with different feature extractors

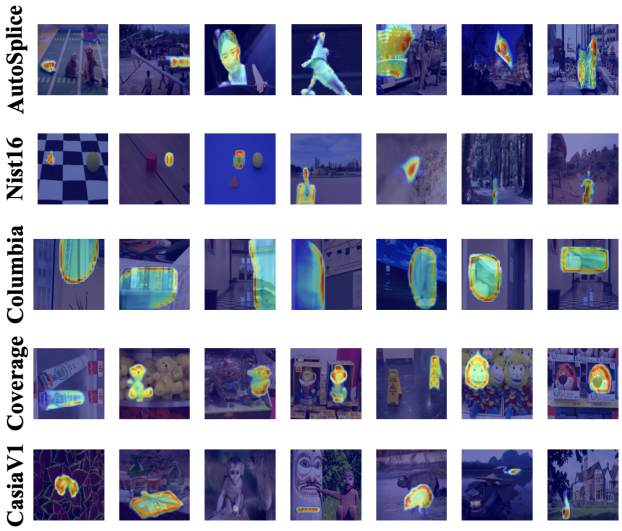

Figure 6: Visualizes the Class Activation Map to highlight the image regions that the models focus on for decision-making.

Performance Comparisons

As shown in Table 1, our method demonstrates clear superiority, improving upon the best baseline by 8.2%, 5.7%, and 8.5% on the first three datasets, and surpassing the suboptimal method, Mesorch, by an average of 3.35%. Furthermore, qualitative visualizations in Figure 5 and CAMs in Figure 6 confirm the model's strong generalization ability. These results show that our method accurately localizes tampered areas despite significant variations in their size and manipulation technique, effectively utilizing a complementary enhancement mechanism for tampering clues to achieve precise detection.

FLOPs and Parameters

The number of parameters and FLOPs for all measurements were calculated based on a resolution of 512x512 and a batch size of 1.

Model	**Parameters (M)**	**FLOPs (G)**
MVSS-Net	150.528	171.008
PSCC-Net	3.668	376.832
TruFor	68.697	236.544
Mesorch	85.754	124.928
Co-Transformers	174	224

Table 3: Comparison of models based on Parameters and FLOPs

As shown in Table 3 , although our model has a relatively large number of parameters and higher memory consumption, it maintains a favorable inference speed by optimizing the FLOPs. As a result, it achieves strong overall performance. This design effectively balances inference efficiency and model accuracy.

Robustness Testing

In this section, we thoroughly evaluate the robustness of Co-Transformers using four test datasets. Following the method proposed by Mesorch, we applied four common attack methods—JPEG compression, Gaussian blur, Gaussian noise, and gamma correction—to generate attacked images under different levels of perturbation. The results are displayed in Figure 7.

The experimental results on the NIST16 and Columbia datasets demonstrate that our method exhibits significant advantages over other methods, particularly in terms of F1-score. This superiority is primarily attributed to our model's

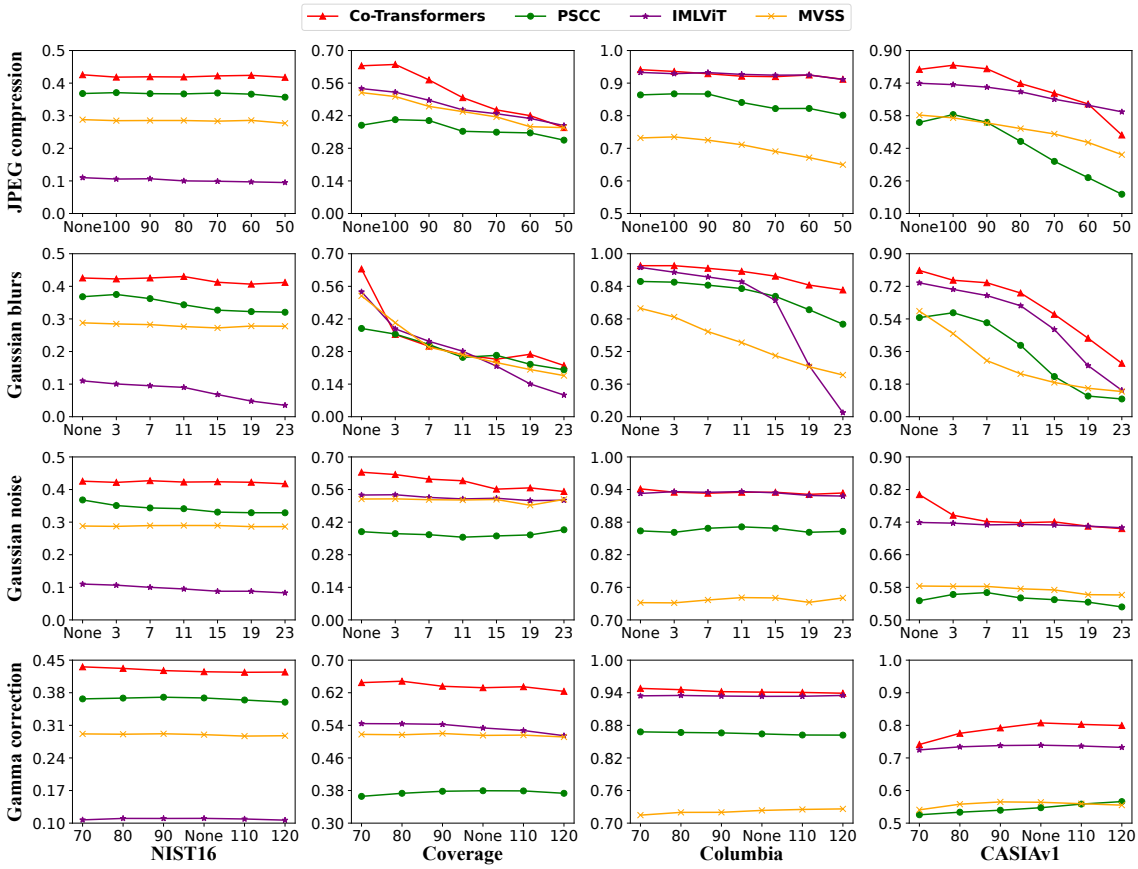

Figure 7: Quantitatively evaluates the models' robustness against various image processing attacks. The plots illustrate the trend of the F1-score as the attack intensity varies.

utilization of collaborative architecture, which enables the hierarchical and comprehensive exploitation of multiple tampering features for detection.

Ablation study

Previous work (Ma et al. 2024b) indicates that feature extractors like BayarConv and Sobel negatively impact model performance. In contrast, extractors such as DCT, FFT, and SRM significantly improve ResNet-based models. However, these same extractors can lead to underfitting in ViT-based models, likely because they require more training epochs to converge. Consequently, the researchers recommend exploring more advanced feature fusion techniques for ViT frameworks.

Building on this insight, we compare the performance of our Co-Transformer architecture, equipped with different combinations of noise extractors, across five benchmark datasets. The results, summarized in Table 2, clearly indicate that our use feature extractor—designed to fuse three noise extraction methods—consistently outperforms all other configurations across multiple test datastes. We attribute this improvement to the task-specific design of our model, which contrasts with previous studies that largely focused on appending noise extractors to standard backbone architectures without tailoring them to the characteristics of IML tasks.

Conclusion

In this paper, we propose a novel Collaborative Transformers framework. This framework achieves complementary enhancement of multi-scale tampering clues through a dual-transformer collaborative mechanism, addressing the problem of insufficient utilization of global semantics and local traces in image manipulation localization. Furthermore, to better utilize multiple noises, we introduce Multi-Level Forensic Attention mechanism. By hierarchically and differentially focusing on the representation of local tampering traces under different receptive fields, it efficiently utilizes higher information density and wider coverage, capturing tampering clues that are difficult to detect with single noise features. Extensive experiments on five test datasets demonstrate that Co-Transformers performs exceptionally well, with better generalization ability and higher robustness, achieving state-of-the-art performance.

Acknowledgments

This work is supported by "National Key Research and Development Program of China" [2024YFF0907404]

References

Dong, C.; Chen, X.; Hu, R.; Cao, J.; and Li, X. 2022. Mvss-net: Multi-view multi-scale supervised networks for image manipulation detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3): 3539–3553.

Dong, J.; Wang, W.; and Tan, T. 2013. Casia image tampering detection evaluation database. In *Proceedings of IEEE China Summit and International Conference on Signal and Information Processing*, 422–426.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *Proceedings of International Conference on Learning Representations*.

Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; and Lu, H. 2019. Dual attention network for scene segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3146–3154.

Guillaro, F.; Cozzolino, D.; Sud, A.; Dufour, N.; and Verdoliva, L. 2023. Trufor: Leveraging all-round clues for trustworthy image forgery detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20606–20615.

Hendrycks, D.; and Gimpel, K. 2016. Gaussian Error Linear Units (GELUs). *arXiv:1606.08415*.

Kwon, M.-J.; Nam, S.-H.; Yu, I.-J.; Lee, H.-K.; and Kim, C. 2022. Learning JPEG compression artifacts for image manipulation detection and localization. *International Journal of Computer Vision*, 130(8): 1875–1895.

Li, S.; Ma, W.; Guo, J.; Xu, S.; Li, B.; and Zhang, X. 2024. UnionFormer: Unified-Learning Transformer with Multi-View Representation for Image Manipulation Detection and Localization. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12523–12533.

Liu, X.; Liu, Y.; Chen, J.; and Liu, X. 2022. PSCC-Net: Progressive spatio-channel correlation network for image manipulation detection and localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(11): 7505–7517.

Loshchilov, I.; and Hutter, F. 2017. SGDR: Stochastic Gradient Descent with Warm Restarts. In *Proceedings of International Conference on Learning Representations*.

Ma, X.; Du, B.; Jiang, Z.; Hammadi, A. Y. A.; and Zhou, J. 2024a. IML-ViT: Benchmarking Image Manipulation Localization by Vision Transformer. *arXiv:2307.14863*.

Ma, X.; Zhu, X.; Su, L.; Du, B.; Jiang, Z.; Tong, B.; Lei, Z.; Yang, X.; Pun, C.-M.; Lv, J.; et al. 2024b. Imdl-benco: A comprehensive benchmark and codebase for image manipulation detection & localization. In *Proceedings of Neural Information Processing Systems*, 134591–134613.

Milletari, F.; Navab, N.; and Ahmadi, S.-A. 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *Proceedings of the 4th International Conference on 3D Vision*, 565–571.

Qu, C.; Zhong, Y.; Liu, C.; Xu, G.; Peng, D.; Guo, F.; and Jin, L. 2024. Towards Modern Image Manipulation Localization: A Large-Scale Dataset and Novel Methods. In *Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10781–10790.

Tan, M.; and Le, Q. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *Proceedings of International conference on machine learning*, 6105–6114.

Triaridis, K.; and Mezaris, V. 2024. Exploring Multi-modal Fusion for Image Manipulation Detection and Localization. In *Proceedings of the 30th International Conference on MultiMedia Modeling*, 198–211.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Proceedings of Neural Information Processing Systems*, 6000–6010.

Wang, J.; Wu, Z.; Chen, J.; Han, X.; Shrivastava, A.; Lim, S.-N.; and Jiang, Y.-G. 2022a. Objectformer for image manipulation detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2364–2373.

Wang, M.; Fu, X.; Liu, J.; and Zha, Z.-J. 2022b. Jpeg compression-aware image forgery localization. In *Proceedings of the 30th ACM International Conference on Multimedia*, 5871–5879.

Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J. M.; and Luo, P. 2021. SegFormer: Simple and efficient design for semantic segmentation with transformers. In *Proceedings of Neural Information Processing Systems*, 12077–12090.

Zhang, Y.; Zhu, G.; Wang, X.; Luo, X.; Zhou, Y.; Zhang, H.; and Wu, L. 2023. CNN-Transformer Based Generative Adversarial Network for Copy-Move Source/ Target Distinguishment. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(5): 2019–2032.

Zhu, X.; Ma, X.; Su, L.; Jiang, Z.; Du, B.; Wang, X.; Lei, Z.; Feng, W.; Pun, C.-M.; and Zhou, J.-Z. 2025. Mesoscopic insights: Orchestrating multi-scale & hybrid architecture for image manipulation localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 11022–11030.